

A CONSTRUÇÃO DE MÁQUINAS FALANTES COMO LUGAR DE INTEGRAÇÃO ENTRE CIÊNCIAS E TECNOLOGIAS DE FALA

ABSTRACT: *This work supports the view of an integration between speech sciences and technologies by the building of a talking machine. Speech production modeling is sketched within speech science framework with special reference to speech synthesis research. The implications of this area of research for teaching are also discussed.*

KEYWORDS: *speech synthesis; speech production; modeling; speech science teaching*

0. Introdução

Quando se aborda a área de pesquisa da Comunicação Falada, normalmente se associam à mesma cinco pólos ou sub-áreas de investigação *à part entière*: Cognição, Produção de Fala, Percepção de Fala, Fonética Acústica e Estudos da Língua. Todas essas sub-áreas em conjunto têm o propósito de esclarecer como se dá o processo comunicativo.

Naquilo que diz respeito à linguagem falada, os estudos cognitivos se concentram no detalhamento dos processos cerebrais que manipulam o conhecimento da língua (ou das línguas), do canal de comunicação, da situação, dos interlocutores e do mundo para elaborar uma mensagem pré-verbal com destinação de um receptor ou para interpretar uma mensagem recebida de um emissor.

A mensagem pré-verbal constitui a primeira etapa do mecanismo de produção de fala, que é seguida das etapas de codificação (a língua é o código sendo manipulado) constituídas pelos processamentos sintático, fonológico e fonético, todas elas se utilizando de conhecimento processual (regras) e representacional (o léxico, por exemplo). As especificações fonéticas para a articulação são transduzidas em comandos neuro-motores que controlam o movimento dos articuladores da fala (vide Levelt, 1989 para uma introdução à área de produção de fala).

A mensagem que se destina a ser interpretada passa por um mecanismo de decodificação (*parsing*), a partir do sistema periférico auditivo e vias auditivas superiores até as áreas corticais de compreensão da fala (área de Wernicke e áreas integrativas).

À Fonética Acústica cabe o estudo das propriedades dos sons da fala. Com o advento de aparelhos para análise do som, como o quimógrafo, no século XIX, e o espectrógrafo, na segunda metade do século XX, além do advento da computação digital, passa a ser possível uma Fonética Instrumental, que passou a trabalhar com a gravação de trechos de som de fala em meio elétrico-eletrônico, constituindo um sinal de fala.

O estudo da língua enquanto sistema de regras e representações cabe à Lingüística, ciência que evidentemente se debruça sobre todas as sub-áreas acima citadas.

Essas mesmas sub-áreas elaboram teorias para explicar o funcionamento de seus objetos de estudo ou desse objeto multifacetado que é a fala. Essas teorias descrevem e manipulam símbolos e escopos de atuação desses mesmos símbolos (com diferentes graus de discretização em cada nível). Para realizar sua tarefa, cada teoria propõe um modelo (normalmente lógico-matemático) que simula, de maneira simplificada em algum sentido, o funcionamento do mecanismo de interesse. A Síntese de Fala se insere nessa perspectiva pois, desde os tempos mais remotos, é um modelo exemplar para a integração de conhecimento científico e tecnológico para a explicitação da produção da fala.

1. Máquinas falantes do passado: simulação da produção de som

O século XVIII é caracterizado por um interesse marcado pela reprodução de autômatos que simulavam com maior ou menor precisão os movimentos naturais. Aclamados por sua perfeição técnica, as três figuras de Jacques de Vaucanson, o pato “digeridor”, o tocador de flauta transversal e o tocador de gaita (Doyon & Liaigre, 1967) alimentaram notavelmente as discussões filosóficas em torno da noção cartesiana de animal-máquina (vide *****, no prelo para uma revisão aprofundada sobre a questão e a história da Síntese de Fala). Na esteira de Vaucanson, um outro construtor de autômatos, o barão von Kempelen, dedicou vinte anos de sua vida à construção de uma máquina falante, manualmente operada e descrita detalhadamente em seu livro, lançado em Viena, em 1791 (Kempelen 1791). Essa máquina (exposta no Deutsches Museum de

Munique e operacional até hoje) foi reproduzida várias vezes no século XIX (por figuras como Sir Charles Wheatstone e Alexander Graham Bell), assinalando o interesse em conhecer o mecanismo de produção do som a partir de um modelo mecânico. O interesse continuará no século XX, com o Voder de Dudley, Riesz e Watkins (1939), uma máquina elétrica que, aproveitando da analogia entre Leis da Mecânica e da Eletricidade (analogia sugerida provavelmente por Graham Bell, a partir da invenção do telefone), simula o funcionamento do aparelho fonador. Tanto elétricas quanto mecânicas, essas máquinas são sintetizadores articulatórios, pois produzem som a partir de modelos dos articuladores da fala.

Sintetizadores acústicos serão possíveis a partir da invenção do gravador operando com fita magnética, ao final do século XIX, e do espectrógrafo (analisador espectral), durante a Segunda Grande Guerra. Através deles será possível manipular diretamente um registro magnético (edição de fitas magnéticas) ou gráfico do sinal de fala (espectrogramas reais ou estilizados) e obter fala sintética.

Nenhum desses dispositivos, no entanto, simula o mecanismo completo da produção de fala, mas tão somente o da produção de som. Para realizar uma simulação completa é preciso integrar o tratamento da informação lingüística e realizar uma passagem não evidente: de uma descrição discreta ou discretizável - do que parece ser a maneira de operar do mundo mental - para uma descrição do contínuo, do mundo físico. Ao integrar a manipulação de informação lingüística, os sintetizadores vão dar lugar aos modernos sistemas de síntese da fala.

2. Máquinas falantes do presente e para o futuro: simulação do mecanismo de produção de fala

Segundo o tipo de simulação do mecanismo de produção de fala que um sistema de síntese realiza, tem-se duas classes de sistemas. Se a simulação parte da elaboração da mensagem pré-verbal, o sistema é de Síntese de Fala a partir do Conceito (CTS, de Concept-to-Speech). Se a simulação parte do texto escrito, simulando a leitura em voz alta, tem-se a Síntese de Fala a partir do Texto (TTS, de Text-to-Speech). Para uma introdução ao primeiro, ver o artigo de Young & Fallside (1979). Para uma introdução ao segundo, ver a excelente revisão de Klatt (1987). Ambos os sistemas possuem módulos de processamento lingüístico que

constituem a etapa mais importante dos mesmos. O processamento é eminentemente sintático-semântico nos sistemas CTS e ortográfico-fônico, nos sistemas TTS. À saída desses sistemas, um sintetizador acústico ou articulatório reproduz o enunciado sintético a partir da seqüência de parâmetros fonético-acústicos (tanto segmentais quanto prosódicos) obtidos a partir da passagem de uma descrição abstrata para uma descrição em termos de números reais (que são uma abstração mais próxima do mundo físico). Um sintetizador acústico pode ser ainda classificado em paramétrico ou concatenativo. O primeiro tipo se serve de especificações da evolução dos parâmetros acústicos como os valores de frequências e larguras de banda de formantes, os valores de frequência fundamental, amplitude, entre outros, para gerar som. O sintetizador concatenativo, por sua vez, opera com um conjunto de trechos sonoros mínimos pré-gravados que são recuperados e têm seus parâmetros prosódicos modificados quando da produção sonora. Esses trechos sonoros, chamados genericamente de polifones, preservam sempre a transição entre duas ou mais realizações consecutivas de fonemas, condição *sine qua non* para a garantia de inteligibilidade (Harris, 1953 e Peterson, Wang & Sivertsen, 1958). Para sintetizar, por exemplo, a palavra “bola” com silêncios inicial e final (grafados aqui com ‘_’) é necessário concatenar os polifones “_b”, “bo”, “ol”, “la” e “a_”.

A melhor maneira de compreender como funciona um sistema de síntese é acompanhando passo a passo o processamento de uma frase escrita qualquer. Usaremos para isso um sistema de síntese TTS genérico.

3. Sintetizando um enunciado em português

Suponhamos que a frase escrita “Há 20 minutos a Sra. Pereira seca a roupa: deve haver um problema na secadora.” seja apresentada ao sistema da figura 1. A primeira etapa é realizada pelo módulo de pré-processamento.

Esse módulo identifica os trechos de texto escrito que não são imediatamente pronunciáveis e realiza uma transformação através de um conjunto de regras de reescrita. Nesse exemplo, ele toma o número “20” e o reescreve por extenso: “vinte”. Em seguida reconhece “Sra.” como uma abreviatura, eliminando a possibilidade de que o ponto seja marca de final de frase e a reescreve por extenso: “senhora” (ou diretamente para uma representação fônica como /sejɔra/). Os sinais de dois pontos e ponto

final podem ser reescritos sob a forma de um código especial para um tratamento ulterior (prosódico, essencialmente).

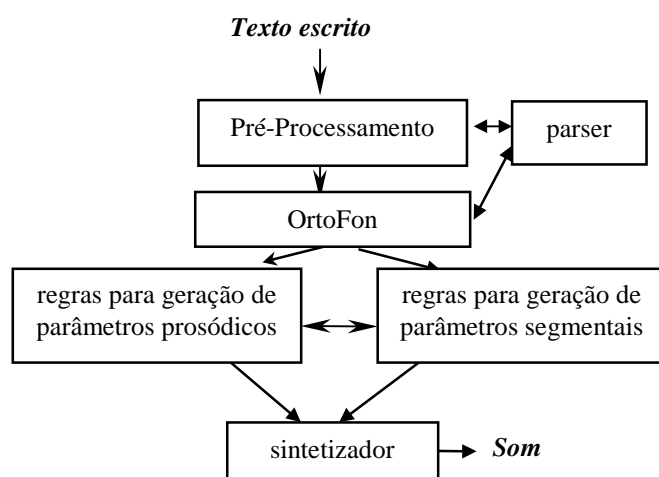


Figura 1: Diagrama geral de um sistema de síntese de fala TTS

O módulo OrtoFon é o responsável pela transdução ortográfico-fônica, também implementado por um conjunto de regras de reescrita (na conhecida forma $A \rightarrow B/C_D$: símbolo A – um grafema – se reescreve B – um fone – se os contextos grafemáticos esquerdo C e direito D estiverem presentes). A frase acima poderá ser reescrita numa representação fonológica como a seguinte: /a viNte minutos a seŋora perejra seka a rowpa deve aver uN problema na sekadora/. Para realizar a transdução do “e” de “seca” no fonema /ε/, o OrtoFon consulta informação morfossintática de um módulo especializado: o parser (nesse caso, “seca” é identificado como terceira pessoa do singular do verbo “secar” e assim pode receber o grau de abertura adequado para a vogal média anterior. O sufixo “-ora”, em “secadora” também é identificado e o /o/ recebe o grau de abertura fechado, simplesmente reescrevendo o sufixo como /ora/).

Para um sistema com sintetizador concatenativo, basta recuperar os polifones apropriados e modificar os parâmetros prosódicos de duração e frequência fundamental dos mesmos para os valores requeridos pela frase (a partir das regras presentes no módulo prosódico, especializado para esse fim). Para um sistema com sintetizador paramétrico, todos os valores numéricos dos parâmetros fonético-acústicos segmentais e prosódicos são fornecidos a uma determinada taxa (a cada 10 ms por exemplo). Ambos os sintetizadores fornecerão a saída sonora ao final desse processo. Exemplos de frases sintéticas obtidas a partir de um sistema de síntese concatenativo do português brasileiro (***** et al. 1999) podem ser avaliados no endereço <<http://www.lafape.iel.unicamp.br>>.

Tendo em vista que durante a produção da fala se dá uma interação entre níveis prosódicos e segmentais, é desejável que as regras operando sobre segmentos e aquelas operando sobre domínios prosódicos possam interagir. É claro que isso só é possível para os sintetizadores articulatório (exigindo nesse caso uma Prosódia Articulatória) e paramétrico. Para o sintetizador concatenativo, a interação deve ser embutida no sinal a ser concatenado, através da gravação prévia de polifones sob diversas condições prosódicas (como em Campbell & Black 1997).

A realização de um sistema de Síntese de Fala exige, evidentemente, a conjugação de uma série de conhecimentos que envolvem as áreas de Ciências da Cognição, Produção e Percepção de Fala, Fonética, Fonologia, Sintaxe, Semântica, Pragmática (para Sistemas Automáticos de Diálogo), Lingüística Computacional, Cálculo, Teoria de Probabilidade, Processamento de Sinal, Engenharia de Software, Estatística, entre outras. É claro que essa constatação coloca questões para a formação de um profissional adequado para colaborar na construção de uma máquina falante: tanto o lingüista quanto o engenheiro de telecomunicações deverão enriquecer sua formação para dar conta do conhecimento lacunar.

4. Algumas implicações para o ensino em Lingüística e em Engenharia de Telecomunicações

A Pós-Graduação em áreas de formação complementares é evidentemente um caminho natural para a formação pluridisciplinar que

exige o trabalho em Ciências da Fala, especialmente a Síntese de Fala. Alguns programas, como o do Instituto de Estudos da Linguagem, na Unicamp e a Faculdade de Engenharia Elétrica da UFMG possuem cursos específicos em Ciência e Tecnologia de Fala. Por abordar a área de um ponto de vista pluridisciplinar, começar por esse tipo de caminho é mais fácil do que a imersão completa numa Pós-Graduação em Lingüística ou em Engenharia de Telecomunicações, respectivamente sendo engenheiro ou lingüista. A partir de um curso pluridisciplinar pode-se migrar para o aprofundamento no campo de mais interesse e necessidade para a pesquisa ou trabalho que se vai desenvolver.

Mas as diferenças entre a maneira de se conduzir a pesquisa (modos de teorizar, métodos empregados e modelos desenvolvidos) nas chamadas Ciências Naturais e nas chamadas Ciências Humanas é normalmente tão grande que um investimento na formação pluridisciplinar desde a Graduação é mais do que desejável, é necessário. O Instituto de Estudos da Linguagem, na Unicamp, tem feito a experiência a partir de 1999, no Bacharelado em Lingüística, com a introdução de um curso que ensina Estatística e Lógica a partir de dados de língua. Resta ainda a avaliar o impacto dessa iniciativa nos anos que se seguirão.

5. Agradecimentos

Ao CNPq (Bolsa de Produtividade em Pesquisa n° 350382/98-0, vinculada ao projeto de número 524110/96-4), e à FAPESP, pelo Auxílio-Pesquisa *Jovem Pesquisador em Centro Emergente* n° 95/09708-6. Esse trabalho também está associado ao Projeto Temático “Integrando Parâmetros Contínuos e Discretos em Modelos do Conhecimento Fônico e Lexical”, n° 01/00136-2, ainda em julgamento.

REFERÊNCIAS BIBLIOGRÁFICAS

- *****, ***** Máquinas Falantes como Instrumentos Lingüísticos: por um Humanismo Éclairé. *Línguas e Instrumentos Lingüísticos*, 8, no prelo.
- _____, VIOLARO, Fábio, ALBANO, Eleonora, SIMÕES, Flávio, AQUINO, Patrícia, MADUREIRA, Sandra, FRANÇOZO, Edson. Aiuruetê: a High-Quality Concatenative Text-to-Speech System for

- Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythm Production. *Proceedings of the Sixth European Conference on Speech Communication and Technology*, Budapeste, Hungria, Setembro 5-9, v 5, 2059-2062, 1999.
- CAMPBELL, Nick W., BLACK, Alan. Prosody and the Selection of Source units for Concatenative Synthesis. In: van SANTEN, J.P.H., SPROAT, R.W., OLIVE, J.P., HIRSCHBERG, J. (Eds.), *Progress in Speech Synthesis*, Nova York: Springer-Verlag, p. 279-292, 1997.
- DOYON, André, LIAIGRE, Lucien. *Jacques Vaucanson, mécanicien de génie*. Paris: Presses Universitaires de France, 1967.
- DUDLEY, Homer, RIESZ, R. R., WATKINS, S. S. A. A synthetic speaker. *Journal of the Franklin Institute*, 227, 739-764, 1939.
- HARRIS, Cyril M. A Study of the Building Blocks in Speech. *J. Acoust. Soc. Am.* 25, 962-969, 1953.
- KEMPELEN, Wolfgang von. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. Vienna: Degen, J. V. (Ed.), [1791] 1970.
- KLATT, Dennis H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.* 82 (3), 737-793, 1987.
- LEVELT, W. *Speaking: from intention to articulation*. Cambridge: MIT Press, 1989.
- PETERSON, G. E., WANG, W. S. Y., SIVERTSEN, E. Segmentation techniques in speech synthesis. *J. Acoust. Soc. Am.* 30 (8), 739-742, 1958.
- YOUNG, S. J., FALLSIDE, F. Speech synthesis from concept: a method for speech output from information systems. *J. Acoust. Soc. Am.* 66 (3), 685-695, 1979.